(19) RÉPUBLIQUE FRANÇAISE

INSTITUT NATIONAL
DE LA PROPRIÉTÉ INDUSTRIELLE

PARIS

(11) N° de publication : 2 763 715

(à n'utiliser que pour les commandes de reproduction)

(21) N° d'enregistrement national : 97 06034

(51) Int Cl⁶ : G 06 F 17/28

Prior Art References Cited in the German Examination Procedure

15

47

Bender, W. et al.: "Techniques for data hiding"; IBM Systems Journal ... attached to the International Search Report

1

(12) **DEMANDE DE BREVET D'INVENTION**

(22) **Date de dépôt :** 22.05.97.

(30) **Priorité :**

(43) **Date de mise à la disposition du public de la demande :** 27.11.98 Bulletin 98/48.

(56) **Liste des documents cités dans le rapport de recherche préliminaire :** *Ce dernier n'a pas été établi à la date de publication de la demande.*

(60) **Références à d'autres documents nationaux apparentés :**

(71) **Demandeur(s) :** *BERTIN ET CIE SOCIETE ANONYME* — FR.

(72) **Inventeur(s) :** MARTEAU PIERRE FRANCOIS et ZNATY ELIE.

(73) **Titulaire(s) :**

(74) **Mandataire(s) :** CABINET ORES.

(54) **PROCEDE DE TRAITEMENT ET DE RECHERCHE D'INFORMATIONS DANS DES DOCUMENTS ENREGISTRES DANS UN SYSTEME INFORMATIQUE.**

(57) Procédé de traitement et de recherche d'informations dans des documents enregistrés dans un système informatique, consistant à indexer les documents pour organiser les termes qui les composent en classes de synonymie associées à des concepts, à établir deux ensembles de règles simples de sémantique et d'association conceptuelle respectivement, et à les appliquer successivement aux documents indexés en spécifiant, pour chaque application, une valeur minimale souhaitée de similarité entre une requête de recherche et les documents indexés.

L'invention réduit les coûts et les temps de calcul des recherches et améliore leur exhaustivité et leur précision.

FR 2 763 715 - A1

PROCEDE DE TRAITEMENT ET DE RECHERCHE D'INFORMATIONS DANS DES DOCUMENTS ENREGISTRES DANS UN SYSTEME INFORMATIQUE.

L'invention concerne un procédé de traitement et de recherche d'informations dans des documents enregistrés dans un système informatique, ce procédé consistant à rédiger une requête de recherche et à l'appliquer aux documents précités au moyen de règles pré-établies pour obtenir les informations recherchées.

Des méthodes informatiques de deux types différents sont actuellement utilisées pour le traitement et l'extraction de l'information documentaire, les unes étant du type numérique et utilisant des moyens statistiques d'analyse, les autres étant du type symbolique et basées sur des moyens de modélisation des connaissances empruntés aux techniques de l'intelligence artificielle.

Ces deux types de méthodes sont complémentaires, car les approches statistiques permettent de couvrir un large domaine à moindre coût avec des capacités de synthèse intéressantes, et les approches symboliques permettent des traitements plus fins sur des domaines de connaissance plus étroits.

L'invention a pour but d'harmoniser et d'intégrer ces deux types de méthodes dans un nouveau procédé de traitement et de recherche d'informations permettant d'améliorer la productivité et les performances et de réduire les coûts au niveau de la modélisation des domaines de connaissance, de l'indexation automatique des documents et de l'extraction automatique des informations recherchées.

Elle propose, à cet effet, un procédé de traitement et de recherche d'informations dans des documents enregistrés dans un système informatique, ce procédé consistant à rédiger une requête de recherche et à appliquer cette requête aux documents précités au moyen de règles pré-établies pour obtenir les informations recherchées, caractérisé en ce qu'il consiste :

- à traiter chaque document par des moyens automatiques d'indexation conceptuelle permettant d'organiser les termes du document en classes de synonymie qui sont associées à des concepts et reliées entre-elles par des relations lexicales et sémantiques,

- à établir des ensembles de règles constituant au moins deux filtres d'informations, le premier composé de règles simples de sémantique et le deuxième composé de règles d'association conceptuelle,

- à définir une fonction de similarité entre un document et une requête,

- et, pour exécution d'une requête donnée,

- à appliquer le premier filtre aux documents indexés en respectant une valeur minimale déterminée de similarité entre les documents et la requête pour obtenir un premier ensemble de documents,

- puis à appliquer le deuxième filtre à cet ensemble de documents en respectant une valeur minimale prédéterminée de similarité entre les documents de cet ensemble et la requête, pour obtenir les informations recherchées.

Le couplage des traitements symboliques et numériques réalisé selon l'invention offre une grande flexibilité au niveau de l'indexation et de l'extraction de l'information, grâce à l'introduction et à la gestion d'une notion de probabilité, liée par exemple aux connaissances incomplètes ou "bruitées".

Ce couplage permet également d'améliorer l'exhaustivité et la précision de la recherche, le filtrage à deux niveaux permettant une simplification de la tâche et une réduction des coûts et des temps de calcul.

Selon une autre caractéristique de l'invention, les valeurs minimales précitées de similarité sont spécifiées par l'utilisateur pour le premier et le deuxième filtre, ou bien sont des valeurs prédéterminées

appliquées automatiquement par le système si l'utilisateur ne spécifie pas de valeurs particulières.

Cette caractéristique de l'invention permet un paramétrage du filtrage par l'utilisateur qui peut ainsi adapter l'exhaustivité et la précision de la recherche à ses besoins particuliers.

Selon encore une autre caractéristique de l'invention, ce procédé consiste également à sélectionner automatiquement celles des règles du deuxième filtre qui sont nécessaires à l'exécution d'une requête donnée, et à n'appliquer que les règles sélectionnées.

On réalise ainsi une adaptation du filtrage à la requête et on réduit les coûts et les temps de calcul.

Selon encore une autre caractéristique de l'invention, on détermine la similarité entre un document et la requête à partir du rapport de la quantité d'information contenue conjointement dans le document et la requête et de la quantité d'information contenue dans la requête.

On peut aussi spécifier une similarité minimale entre deux documents et l'utiliser pour obtenir des classes de documents respectant cette similarité minimale.

De façon générale, l'invention permet d'améliorer les performances et de réduire les coûts du traitement de l'information documentaire, et d'adapter les performances aux besoins ou aux souhaits des utilisateurs.

Elle offre également une plus grande discrétion du traitement, le premier filtre étant par exemple applicable à un grand volume d'informations pour l'obtention d'un ensemble de documents dont la nature n'est pas susceptible de fournir des renseignements à des tiers, le deuxième filtre étant applicable de façon plus confidentielle à cet ensemble de documents déjà extraits du système où ils étaient enregistrés.

L'invention sera mieux comprise et d'autres caractéristiques, détails et avantages de celle-ci apparaîtront plus clairement à la lecture de la description qui suit, faite à titre d'exemple, d'un mode de réalisation particulier de l'invention.

La première phase du procédé selon l'invention comprend une indexation conceptuelle automatique des documents enregistrés, cette indexation consistant à remplacer chaque terme d'un document par un concept tenant compte de liens sémantiques de synonymie, d'hyponymie (spécialisation) ou d'hyperonymie (généralisation).

On peut utiliser à cet effet un système connu de références lexicales, par exemple du type WORDNET pour la langue américaine (une base lexicale développée par l'Université de Princeton), ou EUROWORDNET pour certaines langues européennes dont, à terme, le français et l'allemand, dont la structure s'inspire des théories psycholinguistiques récentes, en particulier des théories sur la mémoire lexicale humaine. Dans un tel système, les noms, les verbes, les adjectifs et les adverbes sont organisés en classes de synonymie que l'on associe à des concepts. Des relations lexicales et sémantiques permettent de lier les classes entres elles, par exemple des relations :

- morphologiques, permettant de spécifier qu'un terme est une forme fléchie d'une racine lexicale,
- antonymiques, permettant de lier des termes contraires (par exemple monter et descendre),
- hyperonymiques ou hyponymiques, permettant d'établir une hiérarchie entre des concepts (par exemple, le terme "couleur" est un concept hyperonyme de "bleu" et, inversement, "bleu" est un concept hyponyme de "couleur"),
- méronymiques ou holonymiques, permettant de spécifier qu'un concept est décomposable en sous-parties

et réciproquement qu'un concept est une sous-partie d'un concept complexe (par exemple "châssis" est un méronyme de véhicule et, inversement, "véhicule" est un holonyme de "châssis").

Ainsi, chaque paragraphe ou chaque phrase d'un texte est traduit dans une séquence de concepts qui constitue une phrase d'un langage conceptuel associé aux moyens sémantiques utilisés. Les règles d'association entre concepts permettent d'enrichir ce langage conceptuel en définissant des concepts plus complexes qui participent également à la phase d'indexation.

On obtient ainsi, à partir d'un document, un ou des fichiers d'index qui associent une liste de références (d'unités de documents) à chacun des termes du document.

Une autre phase du procédé selon l'invention consiste à définir des règles qui vont constituer au moins deux filtres d'information, dont le premier est composé de règles simples de sémantique et le deuxième de règles d'association de concepts.

Les règles du premier filtre sont par exemple des règles de synonymie et d'hyperonymie.

Les règles du deuxième filtre sont des règles d'association et leurs exceptions, permettant de définir une distance (un nombre de mots ou de concepts) et des concepts qui doivent être associés dans cette distance.

Ces règles d'association sont par exemples les suivantes :

- une règle d'association conceptuelle non contrainte, permettant de spécifier que la présence simultanée d'une série de concepts dans la distance D se ré-écrit en un ou plusieurs concepts résultants,
- une règle d'association conceptuelle contrainte, similaire à la règle précédente, à ceci près que l'ordre d'apparition des concepts spécifiés dans la règle doit être respecté,

- des règles d'association terminologique non contrainte et d'association terminologique contrainte, similaires aux deux règles précitées et dans lesquelles seuls les liens de synonymie et d'hyperonomie sont exploités,

- des opérateurs de composition conceptuelle (signes & et @), qui permettent de représenter un concept à partir de plusieurs autres concepts et d'identifier les arguments des premises des règles pour les exploiter dans les conclusions des règles.

Les requêtes établies par les utilisateurs désirant procéder à des recherches sont rédigées en langage naturel ou construites par association de concepts en utilisant des opérateurs du type ET, OU, NON.

L'utilisateur doit également, en principe, spécifier deux degrés de similarité (deux valeurs minimales de similarité) à respecter entre sa requête et les documents recherchés, qui permettent de configurer les bandes passantes des premier et deuxième filtres.

L'invention définit une fonction de similarité entre un document et une requête comme le rapport de la quantité d'information contenue conjointement dans le document et dans la requête et de la quantité d'information contenue dans la requête.

De façon plus détaillée, on peut écrire :

$$P(iu) = \frac{n(iu)}{N}$$

- P(iu) étant la probabilité de trouver une unité d'information (iu) dans un domaine de connaissances,

- n(iu) étant le nombre de documents contenant l'unité d'information (iu) et

- N étant le nombre total d'unités d'information contenues dans ce domaine.

La quantité d'information attachée à l'unité d'information (iu) dans ce domaine est :

$$I(iu) = -\mathrm{Log}_2[P(iu)]$$

La quantité d'information contenue conjointement dans deux documents Di et Dj est :

$$I(Di \cap Dj) = -\sum_{iu} \mathrm{Log}_2\, P(iu)$$

avec $iu \in Di \cap Dj$

La fonction de similarité entre deux documents est alors :

$$S(Di, Dj) = \frac{I(Di \cap Dj)}{\mathrm{Max}[I(Di), I(Dj)]}$$

et la fonction de similarité entre un document Di et une requête R est :

$$S(Di, R) = \frac{I(Di \cap R)}{I(R)}$$

I(R) étant la quantité d'information contenue dans la requête.

La similarité entre deux documents ou entre un document et une requête est un nombre réel compris entre 0 et 1.

Si l'utilisateur fixe une valeur minimale de similarité égale à 0, il aura en réponse à une requête tous les documents d'un domaine de connaissances. S'il fixe une valeur minimale de similarité égale à 1, il n'aura que les documents qui répondent strictement à sa requête.

On demande en principe à l'utilisateur de fixer deux valeurs minimales de similarité, l'une pour l'application du premier filtre et l'autre pour l'application du deuxième filtre.

Si l'utilisateur ne le fait pas, ce sont des valeurs minimales prédéterminées de similarité qui seront appliquées automatiquement par le système.

L'utilisateur ayant formulé une requête et spécifié deux valeurs minimales de similarité pour l'application des deux filtres, le système va d'abord

appliquer le premier filtre (règles de synonymie et d'hyperonymie) aux fichiers d'index constitués à partir des documents faisant partie d'un domaine de connaissance.

Pour cela, le système va prendre le premier terme de la requête et va trouver dans le fichier d'index une liste de références (c'est-à-dire une liste d'unités documentaires).

Le système effectue le rapport du nombre d'unités documentaires de cette liste et du nombre d'unités documentaires dans le domaine de connaissances et obtient une probabilité d'occurrence d'une unité d'information. Le logarithme à base 2 de ce rapport fournit la quantité d'information attachée à cette unité d'information. Ce calcul est fait pour l'ensemble des termes de la requête, ce qui permet d'obtenir la valeur de la similarité entre la requête et la liste de références obtenue. Si cette valeur est supérieure à la valeur minimale spécifiée, la liste de référence est conservée.

L'application des règles de synonymie et d'hyperonymie du premier filtre revient à effectuer ces calculs pour tous les termes du fichier d'index dont les termes de la requête sont des synonymes ou des hyperonymes.

L'application du premier filtre au domaine de connaissance fournit ainsi une ensemble de documents auxquels le deuxième filtre va être appliqué.

Pour cela, le système commence par sélectionner celles des règles du second filtre qui sont nécessaires à l'exécution de la requête et n'applique que les règles ainsi sélectionnées à l'ensemble des documents résultant du premier filtrage.

Les calculs de similarité sont réalisés comme décrit plus haut, en tenant compte des règles d'association conceptuelle qui ont été sélectionnées par le

système, qui modifie les listes d'index associées aux documents sélectionnés à l'issue du premier filtrage.

On obtient ainsi des documents (des unités documentaires) qui répondent à la requête avec une exhaustivité et une précision déterminées par les degrés de similarité spécifiés par l'utilisateur.

On comprend qu'en général le degré de similarité spécifié pour l'application du premier filtre sera relativement faible, pour favoriser l'exhaustivité de la recherche, tandis que celui spécifié pour l'application du second filtre pourra être plus élevé, afin d'augmenter la précision.

Le système permet également à l'utilisateur d'élaborer des règles spécifiques de synonymie, d'hyperonymie, et d'association conceptuelle, qui viendront compléter les règles pré-existantes et qui seront adaptées à la recherche que l'utilisateur souhaite effectuer.

Le procédé selon l'invention permet de faire, non seulement du filtrage et de l'extraction d'informations dans un domaine de connaissances, mais également de fournir des classes documentaires dans lesquelles figurent des documents qui sont sélectionnés à partir de leur similarité (par application de la fonction de similarité entre deux documents qui est indiquée plus haut et comparaison de la similarité à une valeur minimale déterminée ou par application d'algorithme(s) simple(s) de classification automatique du type "nuées dynamiques" qui exploitent la distance d entre deux documents Di, Dj, cette distance étant définie par la relation :

$$d = 1 - S(Di, Dj).$$

Par ailleurs, on peut également considérer le traitement d'indexation automatique des documents comme un premier filtrage ou filtrage préalable et spécifier pour ce filtrage une valeur minimale de similarité. Dans ce cas, on ne retiendra des documents indexés que ceux qui respectent cette valeur minimale de similarité avec

la requête, et les deux autres filtres ne seront appliqués qu'aux documents indexés retenus.

REVENDICATIONS

1) Procédé de traitement et de recherche d'informations dans des documents enregistrés dans un système informatique, ce procédé consistant à rédiger une requête de recherche et à appliquer cette requête aux documents précités au moyen de règles pré-établies pour obtenir les informations recherchées, caractérisé en ce qu'il consiste :

- à traiter chaque document par des moyens automatiques d'indexation conceptuelle permettant d'organiser les termes du document en classes de synonymie qui sont associées à des concepts et reliées entre elles par des relations lexicales et sémantiques,

- à établir des ensembles de règles constituant au moins deux filtres d'informations, le premier composé de règles simples de sémantique et le deuxième composé de règles d'association conceptuelle,

- à définir une fonction de similarité entre un document et une requête,

- et, pour exécution d'une requête donnée,

- à appliquer le premier filtre aux documents indexés en respectant une valeur minimale déterminée de similarité entre les documents et la requête pour obtenir un premier ensemble de documents,

- puis à appliquer le deuxième filtre à cet ensemble de documents en respectant une valeur minimale prédéterminée de similarité entre les documents de cet ensemble et la requête, pour obtenir les informations recherchées.

2) Procédé selon la revendication 1, caractérisé en ce que les valeurs minimales précitées de similarité sont spécifiées par l'utilisateur pour le premier et pour le deuxième filtre, ou bien sont des valeurs prédéterminées appliquées automatiquement par le système si l'utilisateur ne spécifie pas de valeurs particulières.

3) Procédé selon la revendication 1 ou 2, caractérisé en ce que les règles du premier filtre sont des règles de synonymie et d'hyperonymie.

4) Procédé selon l'une des revendications précédentes, caractérisé en ce que les règles du deuxième filtre sont des règles d'association de concepts et leurs exceptions.

5) Procédé selon l'une des revendications précédentes, caractérisé en ce qu'il consiste à sélectionner automatiquement celles des règles du deuxième filtre qui sont nécessaires à l'exécution d'une requête donnée et à n'appliquer que ces règles sélectionnées.

6) Procédé selon l'une des revendications précédentes, caractérisé en ce qu'on détermine la similarité entre un document et une requête à partir du rapport de la quantité d'information contenue conjointement dans ce document et la requête et de la quantité d'information contenue dans la requête.

7) Procédé selon l'une des revendications précédentes, caractérisé en ce qu'il consiste également à définir une fonction de similarité entre deux documents par le rapport de la quantité d'information contenue conjointement dans les deux documents et du maximum des quantités d'information contenues dans les deux documents, et à appliquer cette fonction de similarité aux documents indexés pour obtenir une classification des documents.

8) Procédé selon l'une des revendications précédentes, caractérisé en ce que, lors de l'indexation des documents, on définit pour chaque document un fichier d'index constitué de lemmes qui sont des formes lexicales réduites des mots du document et, pour l'exécution d'une requête, on applique les règles du premier filtre aux fichiers d'index.

9) Procédé selon la revendication 8, caractérisé en ce que les règles du deuxième filtre sont appliquées aux fichiers d'index.

10) Procédé selon l'une des revendications précédentes, caractérisé en ce qu'il consiste à faire rédiger par un utilisateur des règles de synonymie, d'hyperonymie et d'association conceptuelle qui sont spécifiques à une recherche particulière et à prendre ces règles en compte, avec les règles pré-existantes constituant les premier et deuxième filtres pour l'exécution de la recherche.

11) Procédé selon l'une des revendications précédentes, caractérisé en ce qu'il consiste à spécifier une valeur minimale prédéterminée de similarité entre une requête et les documents avant d'effectuer le traitement précité d'indexation conceptuelle des documents, et à ne retenir que ceux des documents indexés qui respectent cette valeur minimale de similarité.